Topics in Computational Biology and Genomics

Plant & Microbial Biology c146
Molecular & Cell Biology c146
Bioengineering c146

Plant & Microbial Biology c246
Molecular & Cell Biology c246
Bioengineering c246

University of California, Berkeley
Spring 2003

"Instruction and discussion of topics in genomics and computational biology. Working from evolutionary concepts, the course will cover principles and application of molecular sequence comparison, genome sequencing & functional annotation, and phylogenetic analysis."

4 Units

Instructors.

Steven E. Brenner, Assistant Professor, Plant & Microbial Biology
Affiliated Assistant Professor, Molecular & Cell Biology, Bioengineering
Faculty Scientist, Lawrence Berkeley National Laboratory
Michael B. Eisen, Staff Scientist, Lawrence Berkeley National Laboratory
Assistant Adjunct Professor, Molecular & Cell Biology
Both may be reached by email to profs@c246.lbl.gov

Teaching assistant.

Derek Chiang, Graduate Student, Molecular & Cell Biology
derek@rana.lbl.gov

Class meetings.

Tuesday and Thursday, 11am-12:30 pm in Warren 24
Weekly discussion section: Friday 9:00-10:00 am in 107 GPBB.
Attendance is required.

Prerequisites.

Bioengineering 142, Computer Science 61B, or equivalent ability to write programs in Java, Perl, C, or C++; Molecular and Cell Biology 100, 102, or equivalent; or consent of instructor

Core specialization (Bioengineering).

B (Bioinformatics and Genomics) and D (Computational Bioengineering). It also fulfills biological content.

Core requirement (Molecular and Cell Biology).

This course fulfills a core course requirement for the department of Molecular and Cell Biology.

Textbook.

Durbin R., Eddy S., Krogh A., Mitchison G. Biological Sequence Analysis. Cambridge: Cambridge UP, 1998.

Literature articles found on the course website: http://c246.lbl.gov

Assigned readings must be completed before the class for which they are assigned.

Optional Additional References.

These books provide additional introductory references to the core topics that will be discussed in the course. Copies will be placed on reserve in the Biosciences Library.

Lesk, A.M. Introduction to Bioinformatics. Oxford: Oxford UP, 2002.

Hall, B.G. Phylogenetic Trees Made Easy. Sinauer Associates, 2001.

Koonin E.V., Galperin M.Y. Sequence – Evolution – Function: Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, 2002.

Setubal J.C., Meidanis J. Introduction to Computational Molecular Biology. Brooks/Cole Pub Co, 1997.

Grading.

25% homework
20% midterm exam
20% project
25% final exam (+ resurrection from midterm exam)
10% class participation

Membership.

The six different "versions" of the class. The versions listed in different departments are *identical*. You may sign up for any version.

The undergraduate c146 & graduate c246 versions have the same lectures. However, for the graduate version, students will be required to do additional questions on homework problem-sets and to prepare a paper presentation for the class section.

Auditors are welcome if space allows. Auditors are expected to participate fully in the class

Homework.

Homework will typically be assigned in class on Tuesdays, and it will be due by email to by 5pm the following Monday. Homework should be submitted to derek@rana.lbl.gov and should be in plaintext, Word or PDF. *Identical* paper copies must be turned in at the beginning of class on Tuesday. Where verbal responses are required, they must be in cogent standard written English.

Oral discussion of the class and homework is encouraged. However, all homework questions must be answered in writing alone and must be fully understood. You must also list all the people with whom you discussed the question.

Homework received between 5pm Monday and 11am Tuesday will be penalized by 10 percent. After that, an additional 10 percent will be deducted for each day late, and no credit will be given for problems that have been discussed in class.

The lowest scoring homework will not be included in your grade calculation.

Computer access.

Programs may be written on any computer in Perl, C, C++, or Java.

Class notes.

For lectures given with PowerPoint, the instructor's presentation will be placed on the course website following class.

This class will use the scribe system. *Failure to adhere to the following requirements will impact the student's class participation grade*. One student ("scribe") will be designated to take notes each week, while another ("reader") will review these notes for accuracy and work with the scribe to correct any errors or omissions. The scribe must provide notes to the reader by the following lecture. By the lecture thereafter, the reader must submit the notes by email to the teaching assistant. All notes must be electronic so they may be placed on the website.

Office hours.

Office hours for Steven Brenner will be 5:30-7:00pm on Mondays in Koshland Hall 461-East; office hours for Michael Eisen will be by appointment. Any changes in office hours will be announced.

Project.

Pairs (or for exceptionally complex projects, triples) of students will undertake a substantial research project, creating new computational biology methodologies or performing a significant genomic analysis. The final project will be presented at a class poster session and written up as a brief (roughly 3 page) report. Electronic versions of both the poster and report must be submitted, along with supplementary information including figures, references, datasets, and custom software.

Website.

The course website is http://c246.lbl.gov. Consult the page regularly for homework, class notes, and updated information.

Course Schedule

January 21
Lecture 1 [Brenner/Eisen]
Introduction to Sequence Analysis

NO READINGS

January 23
Lecture 2 [Brenner]
Sequence Evolution

This lecture will discuss evolution at the sequence level and the importance of understanding evolution for sequence analysis.
Read Walter Fitch's modern discussion of homology and terminology used in discussing sequence evolution. Then read short note by Winter et al. which argued the reconstruction of sequence evolution is impossible, and Fitch's rebuttal.
Begin reading the ISMB tutorial on protein evolution by William Pearson (complete by Lecture 3).
Pearson W.(2000). **Protein sequence comparison and protein evolution** This is the ISMB tutorial.
Fitch WM.(2000). **Homology a personal view on some of the problems**. *Trends in Genetics* 16:227-31.
Winter WP, Walsh KA and Neurath H.(1968). **Homology as applied to proteins**. *Science* 162:1433.
Fitch WM.(1970). **Distinguishing homologous from analogous proteins**. *Systematic Zoology* 19:99-113.For now, just read pages 99-102, 112-113.

January 28
Lecture 3 [Brenner]
Sequence Evolution

Start reading chapter 2 of the Durbin, Eddy, Krogh & Mitchison (DEKM) book.
Finish reading pages 103-111 of the Fitch article. Focus on understanding principles, but not the details.
Continue reading the Pearson ISMB tutorial.

January 30
Lecture 4 [Brenner]
Scoring Alignments; Dynamic Programming

DOTTER, dot plots, local & global alignments

Finish reading DEKM sections 2.1, 2.2, 2.3

February 4
Lecture 5 [Brenner]
Dynamic Progamming with General Gap Penalties
Affine gaps, dynamic programming, gap parameters

DEKM section 2.4

February 6
Lecture 6 [Brenner]
Matrices and Gap Parameters

DEKM section 2.8
Henikoff S and Henikoff JG (1992). **Amino acid substitution matrices from protein blocks**. *Proceedings of the National Academy of Sciences of the United States of America* 89:10915-19.
Dayhoff MO, Schwartz RM and Orcutt BC (1978). **A model of evolutionary change in proteins**.

February 11
Lecture 7 [Brenner]
Heuristic Alignment Methods [FASTA and BLAST]

DEKM section 2.5
Gallison F (2000). **The Fasta and Blast programs**

February 13
Lecture 8 [Brenner]
Statistical Significance of Alignments

DEKM section 2.7

February 18
Lecture 9 [Brenner]
Biological Significance of Alignments

Brenner SE (1999). **Errors in genome annotation**. *Trends in Genetics* 15:132-3.
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al.* (2000). **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 25:25-29.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999). **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles** *Proc Natl Acad Sci USA* 96:4285-4288.
Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D (1999). **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 285:751-753.

February 20
Lecture 10 [Brenner]
Multiple Alignment I
CLUSTALW, T-COFFEE, MAFFT

DEKM Chapter 6
Higgins DG, Thompson JD and Gibson TJ (1996). **Using CLUSTAL for multiple sequence alignment.** *Methods in Enzymology* 266:383-402.
Notredame C, Higgins DG and Heringa J (2000). **T-Coffee: A novel method for dast and accurate multiple sequence alignment.** *Journal of Molecular Biology* 302:205-217.
Katoh K, Misawa K, Kuma K and Miyata T (2002). **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Research* 30:3059-3066.

February 25
Lecture 11 [Brenner]
Multiple Alignment II

DEKM Chapter 6

February 27
Lecture 12 [Brenner]
Multiple Alignment III

DEKM Chapter 6

March 4

MIDTERM EXAM

March 6
Lecture 13 [Brenner]
Practical Database Searching

BLAST, PFAM, CDS, SMART

March 11
Lecture 14 [Brenner]
Practical Database Searching (Iterated)

SCOP, SWISSPROT, PDB, NCBI NR, PSI-BLAST, T99, SUPERFAMILY
Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, et al. (2001). **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Res* 29:2994-3005.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Research* 25:3389-3402.
[SAM-T99 paper]

March 13

Review of Midterm Exam

March 18
Lecture 15 [Eisen]
Finding Motifs I: MEME, EM and Gibbs Sampling

Stormo GD (2000). DNA binding sites: representation and discovery *Bioinformatics* 16:16-23.
Bailey TL and Elkan C (1994). **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.
Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC (1993). **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 262:208-14.

March 20
Lecture 16 [Eisen]
Finding Motifs II: Dictionaries and Applications

Bussemaker HJ, Li H and Siggia ED (2000). **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 97:10096-100.
Hughes JD, Estep PW, Tavazoie S and Church GM (2000). **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae**. *Journal of Molecular Biology* 296:1205-14.
Spring Break: Read about Hidden Markov Models (Durbin chapter 3. pg. 48-71)

April 1
Lecture 17 [Eisen]
What are Hidden Markov Models and Why Use Them?

DEKM Chapter 3

April 3
Lecture 18 [Eisen]
Hidden Markov Models

DEKM Chapter 3

April 8
Lecture 19 [Eisen]
Biological Applications of Hidden Markov Models

DEKM Chapter 5

April 10
Lecture 20 [Eisen]
Generalizations of Hidden Markov Models

Kulp D, Haussler D, Reese M and Eeckman FH (1996). **A generalized Hidden Markov Model for the recognition of human genes in DNA.** *Proc Int Conf Intell Syst Mol Biol* 4:134-142.
Reese MG, Kulp D, Tammana H and Haussler D (2000). **Genie--gene finding in *Drosophila melanogaster*** *Genome Research* 10:529-538.
Burge C and Karlin S (1997). **Prediction of complete gene structures in human genomic DNA** *Journal of Molecular Biology* 268:78094.

April 15
Lecture 21 [Eisen]
Why Phylogeny Matters

Doolittle WF (1999). **Phylogenetic classification and the universal tree**. *Science* 284:2124-2129.
Eisen JA (1998). **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis**. *Genome Research* 8:163-7.

April 17
Lecture 22 [Eisen]
Phylogeny: Distance Methods

DEKM Chapter 7
Page RDM and Holmes EC (1998). **Molecular evolution : a phylogenetic approach**, (Oxford ; Malden, MA, Blackwell Science), pp. 172-227.

April 22
Lecture 23 [Eisen]
Phylogeny: Parsimony

April 24
Lecture 24 [Gavin Crooks]
Phylogeny: Bootstrapping

Felsenstein J (2002). **Bootstrap and randomization tests.** Chapter 20 of **Inferring Phylogenies**, (Cambridge, MA, Sinnauer).
Huelsenbeck JP and Rannala B (1997). **Phylogenetic methods come of age: testing hypotheses in an evolutionary context**. *Science* 276:227-32.

April 29
Lecture 25 [Eisen]
Likelihood

DEKM Chapter 8
Page RDM and Holmes EC (1998). **Molecular evolution : a phylogenetic approach**, (Oxford ; Malden, MA, Blackwell Science), pp. 193-200.
Felsenstein J (2002). **Likelihood methods.**. Chapter 16 of **Inferring Phylogenies**, (Cambridge, MA, Sinnauer).

May 1
Lecture 26 [Eisen]
Practical Likelihood Methods and Combined Alignment and Phylogeny

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001). **Bayesian inference of phylogeny and its impact on evolutionary biology**. *Science* 294:2310-4.
And for the brave
Mitchison GJ (1999). **A probabilistic treatment of phylogeny and sequence alignment**. *Journal of Molecular Evolution* 49:11- 22.
Holmes I and Bruno WJ (2001). **Evolutionary HMMs: a Bayesian approach to multiple alignment**. *Bioinformatics* 17:803-20.

May 6
Lecture 27 [Brenner & Eisen]
All Questions Answered

May 8

Final Project Consultation

May 13

POSTER PRESENTATIONS

May 15

POSTER PRESENTATIONS

FINAL EXAM
Time and location TBA